Faster than FAST: GPU-accelerated frontend for high-speed VIO

Nagy, Balazs ; Foehn, Philipp ; Scaramuzza, Davide

This paper has been accepted for publication at the
IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020. ©IEEE

Faster than FAST: GPU-Accelerated Frontend for High-Speed VIO

Balázs Nagy, Philipp Foehn, Davide Scaramuzza

***Abstract*—The recent introduction of powerful embedded graphics processing units (GPUs) has allowed for unforeseen improvements in real-time computer vision applications. It has enabled algorithms to run onboard, well above the standard video rates, yielding not only higher information processing capability, but also reduced latency. This work focuses on the applicability of efficient low-level, GPU hardware-specific instructions to improve on existing computer vision algorithms in the field of visual-inertial odometry (VIO). While most steps of a VIO pipeline work on visual features, they rely on image data for detection and tracking, of which both steps are well suited for parallelization. Especially non-maxima suppression and the subsequent feature selection are prominent contributors to the overall image processing latency. Our work first revisits the problem of non-maxima suppression for feature detection specifically on GPUs, and proposes a solution that selects local response maxima, imposes spatial feature distribution, and extracts features simultaneously. Our second contribution introduces an enhanced FAST feature detector that applies the aforementioned non-maxima suppression method. Finally, we compare our method to other state-of-the-art CPU and GPU implementations, where we always outperform all of them in feature tracking and detection, resulting in over 1000fps throughput on an embedded Jetson TX2 platform. Additionally, we demonstrate our work integrated into a VIO pipeline achieving a metric state estimation at ∼200fps.**

Code available at: https://github.com/uzh-rpg/vilib

I. Introduction

A. Motivation

As technology became increasingly affordable, vision-based motion tracking has proven its capabilities not only in robotics applications, such as autonomous cars and drones but also in virtual (VR) and augmented (AR) reality and mobile devices. While visual-inertial odometry (VIO) prevails with its low cost, universal applicability, and increasing maturity and robustness, it is still computationally expensive and introduces significant latency. This latency impacts e.g. VR/AR applications by introducing motion sickness, or robotic systems by constraining their control performance. The latter is especially true for aerial vehicles with size and weight constraints limiting the available computation power while requiring real-time execution of the VIO and control pipeline to guarantee stable, robust, and safe operation. Besides latency, one may also witness a disconnect between the available sensor capabilities (both visual and inertial) and the actual information processing capabilities of mobile systems. While off-the-shelf cameras are capable of capturing images above 100fps, many algorithms and implementations are not able to handle visual information at this rate. By lowering the frame processing times, we can simultaneously minimize latency and also reduce the neglected visual-inertial information.

All authors are with the Robotics and Perception Group, Dep. of Informatics, University of Zurich, and Dep. of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland - http://rpg.ifi.uzh.ch. Their work was supported by the SNSF-ERC Starting Grant and the Swiss National Science Foundation through the National Center of Competence in Research (NCCR) Robotics.

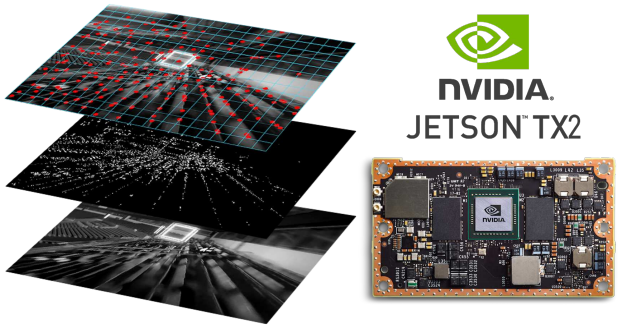


Fig. 1. Our method introduces a novel non-maxima suppression scheme exploiting GPU parallelism and low-level instructions, applied for GPU-optimized feature detection and tracking. We demonstrate a feature detection and tracking rate of **over 1000fps** on an embedded Jetson TX2 platform.

In particular, embedded systems on drones or AR/VR solutions cannot rely on offline computations and therefore need to use all their available resources efficiently. Various heterogeneous embedded solutions were introduced, offering a range of computing architectures for better efficiency. There are three popular heterogeneous architectures: (i) the first one uses a central processing unit (CPU) with a digital signal processor (DSP), and therefore it is restricted in its set of tasks; (ii) the second one combines a CPU with programmable logic (e.g. FPGA), which is versatile but increases development time; (iii) the third solution is the combination of a CPU with a GPU, which is not only cost-efficient but also excels in image processing tasks since GPUs are built for highly parallel tasks.

On these grounds, our work investigates feature detection and tracking on CPU-GPU platforms, where we build up the image processing from the GPU hardware's perspective. We present a feasible work-sharing between the CPU and a GPU to achieve significantly lower overall processing times.

B. Related Work

We recapitulate previous approaches according to the building blocks of a VIO pipeline front-end: feature detection, non-maxima suppression, and feature tracking.

1) Feature Detection: Over the years, feature detection has not changed significantly. Pipelines commonly use Harris [1], Shi-Tomasi [2], or FAST features [3] with one of three

FAST corner scores. Harris and Shi-Tomasi are less sensitive to edges and are also widely used independently as corner detectors. They both share the same principles, but their metric of cornerness differs. ORB [4], as an extension to FAST has also appeared in VIO pipelines, presenting a reasonable, real-time alternative to SIFT [5] and SURF [6]. Amongst the above, undoubtedly FAST presents the fastest feature detector. Two variations were proposed from the original authors in the form of FAST [3] and FAST-ER [7], where the latter greatly improves on repeatability while still maintaining computational efficiency. To the best of our knowledge, the fastest CPU implementation of the FAST detector is KFAST [8], which showcases more than $5\times$ speedup over the original implementation. On the GPU, we are aware of two optimized implementations within OpenCV [9] and ArrayFire [10]. Both employ lookup tables to speed up the decision process for determining a point's validity: the latter uses a 64-kilobyte lookup table, while the former an 8-kilobyte one. Although both solutions provide fast cornerness decision, neither of them guarantees spatial feature distribution as they both stop extracting features once the feature count limit is reached.

2) Non-Maxima Suppression: Non-maxima suppression can be considered a local maximum search within each candidate's Moore neighborhood. The Moore neighborhood of each pixel-response is its square-shaped surrounding with a side length of $(2n+1)$. Suppression of interest point candidates has been studied extensively in [11], [12], and recently, they have also been reviewed in machine learning applications [13]. The complexity of the proposed algorithms is determined based on the number of comparisons required per interest point. This algorithm requires $(2n+1)^2$ comparisons, as the comparisons follow the raster scan order. Förstner and Gülch [14] proposed performing the comparisons in spiral order. Theoretically, the number of comparisons does not change, however, the actual number of comparisons plummeted, because most candidates can be suppressed in a smaller 3x3 neighborhood first, and only a few points remain. Neubeck [11] proposed several algorithms to push down the number of comparisons to almost 1 in the worst-case. Pham [12] proposed another two algorithms (scan-line, and quarter-block partitioning) that drove down the number of comparisons below 2, not only for larger ($n \geq 5$) but also for small neighborhood sizes ($n < 5$). All these approaches first try to perform reduction to a local maximum candidate in a smaller neighborhood, then perform neighborhood verification for only the selected candidates. However, they do not ensure any (spatial) feature distribution and possibly output a large set of feature candidates.

3) Feature tracking: Feature tracking may be divided into three different categories: (i) feature matching, (ii) filter-based tracking, and (iii) differential tracking. Feature matching (i) applies feature extraction on each frame followed by feature matching, which entails a significant overhead. Moreover, the repeatability of the feature detector may adversely influence its robustness. However, there are well-known pipelines, opting for this approach [15], [16], [17]. [18] tracks the features using filters (ii), which contain the feature location in its state (e.g. via bearing vectors), and follows features with consecutive prediction and update steps. The third differential (iii) approaches aim to directly use the pixel intensities and minimize a variation of the photometric error. From the latter kind, the Lucas-Kanade tracker [19], [20], [21] became ubiquitous in VIO pipelines [22], [23] due to its efficiency and robustness. As it directly operates on pixel intensity patches, GPU adaptations appeared early on. [24] implements a translational displacement model on the GPU with intensity-gain estimation. [25], [26] go even further: they propose an affine-photometric model coupled with an inertial measurement unit (IMU) initialization scheme: the displacement model follows an affine transformation, in which the parameters are affected by the IMU measurements between consecutive frames, while the pixel intensities may also undergo an affine transformation.

C. Contributions

Our work introduces a novel non-maxima suppression building upon [14] but also exploiting low-level GPU instruction primitives, completed by a GPU-optimized implementation of the FAST detector with multiple scores. Our method combines the feature detection and non-maxima suppression, guaranteeing uniform feature distribution over the whole image, which other approaches need to perform in an extra step. Additionally, we combine our frontend with a state-of-the-art VIO bundle-adjustment backend. All contributions are verified and thoroughly evaluated using the EuRoC dataset [27] on the Jetson TX2 embedded platform and a laptop GPU. Throughput capabilities of over 1000fps are demonstrated for feature detection and tracking, and $\sim$200fps for a VIO pipeline recovering a metric state estimate.

II. Methodology

A. Preliminaries on parallelization

We briefly introduce the fundamentals of the Compute Unified Device Architecture, or shortly CUDA, which is a parallel computing platform and programming model proprietary to NVIDIA. CUDA allows developers to offload the central processing unit, and propagate tasks to the GPU even for non-image related computations. Latter is commonly referred to as general-purpose GPU (GPGPU) programming.

The NVIDIA GPU architecture is built around a scalable array of multithreaded streaming multiprocessors (SMs) [28]. Each SM has numerous streaming processors (SP), that are lately also called CUDA cores. GPUs usually have 1-20 streaming multiprocessors and 128-256 streaming processors per SM. In addition to the processing cores, there are various types of memories available (ordered by proximity to the processing cores): register file, shared memory, various caches, off-chip device, and host memory.

NVIDIA's GPGPU execution model introduces a hierarchy of computing units: threads, warps, thread blocks, and thread grids. The smallest unit of execution is a thread. Threads are grouped into warps: each warp consists of 32 threads. Warps are further grouped into thread blocks. One thread block is guaranteed to be executed on the same SM. Lastly, on top of

TABLE I
MEMORY SELECTION FOR THE FASTEST COMMUNICATION

Execution Unit	Execution Unit	Fastest Memory
Threads within warp	identical SM	registers
Warps within thread block	identical SM	shared memory
Thread blocks	any SM	global memory

the execution model is the thread grid. A thread grid is an array of thread blocks. Thread blocks within a thread grid are executed independently from each other.

The instruction execution on the GPU needs to be emphasized: every thread in a warp executes the same instruction in a lock-step basis. NVIDIA calls this execution model Single Instruction Multiple Threads (SIMT). It also entails, that *if/else* divergence within a warp causes serialized execution.

The underlying GPU hardware occasionally undergoes significant revisions, hence the differences between GPUs need to be tracked. NVIDIA introduced the notion of Compute Capability accompanied by a codename to denote these differences. With the introduction of the NVIDIA Kepler GPU microarchitecture, threads within the same warp can read from each other's registers with specific instructions. Our work focuses on these warp-level primitives, more specifically, highly-efficient communication patterns for sharing data between threads in the same warp. In previous GPU generations, threads needed to turn to a slower common memory (usually the shared memory) for data sharing, which resulted in significant execution latencies. However, with the introduction of the Kepler architecture it became possible to perform communication within the warp first, and only use the slower memory on higher abstractions of execution, i.e. within thread blocks and then within the thread grid. In Table I we summarized the available fastest memories for exchanging data between execution blocks on a GPU.

B. Feature detector overview

GPUs are particularly well-suited for feature detection, which can be considered a stencil operation amongst the parallel communication patterns. In a stencil operation, each computational unit accesses an input element (e.g. pixel) and its close neighborhood in parallel. Therefore, the image can be efficiently divided amongst the available CUDA cores, such that the memory accesses are coalesced, leading to highly effective parallelization. For feature detection, the input image is first subsampled to acquire an image pyramid. Then, for each image resolution, two functions are usually evaluated at each pixel: a coarse corner response function (CCRF), and a corner response function (CRF). CCRF serves as a fast evaluation that can swiftly exclude the majority of candidates so that a slower CRF function only receives candidates that passed the first verification. Once every pixel has been evaluated within the ROI, non-maxima suppression is applied to select only the local maxima. We summarized the general execution scheme of feature detector algorithms in Algorithm 1. It was discovered in [29], [30], [31], that uniform feature distribution

Algorithm 1: Generalized Feature Detection

```
for Every scale do
    for Every pixel within the region of interest (ROI) do
        if Coarse Corner Response Function (CCRF) then
            Corner Response Function (CRF)
        end
    end
    Non-max. suppression within neighborhood (NMS)
end
• Non-max. suppression within cell (NMS-C)
```

on image frames improves the stability of VIO pipelines. To fulfill this requirement, [22] and [23] introduced the notion of 2D grid cells: the image is divided into rectangles with a fixed width and height. Within each cell, there is only one feature selected - the feature whose CRF score is the highest within the cell. Not only does this method distribute the features evenly on the image, but it also imposes an upper limit on the extracted feature count. This is shown in Algorithm 1, including the augmentation •.

C. Non-maxima suppression with CUDA

Feature selection within a cell can be understood as a reduction operation, where only the feature with the maximum score is selected. Moreover, non-maxima suppression within the neighborhood can also be considered a reduction operation, where one reduces the corner response to a single-pixel location within finite neighborhoods.

Our approach divides the corner response map into a regular cell grid. Within the grid on the first pyramid level, cells use a width of an integer multiple of 32, i.e. $32w$, because, on NVIDIA GPU hardware, a warp consists of 32 threads. One line of a cell is referred to as a cell line, which can be split up into cell line segments with 32 elements. We restrict the height of the cells to be $2^{l-1}h$, where l is the number of pyramid levels utilized during feature detection. By selecting an appropriate grid configuration l, w, and h, one can determine the maximum number of features extracted, while maintaining spatial distribution.

Within one cell line segment, one thread is assigned to process one pixel-response, i.e. one warp processes one entire cell line segment (32 threads for 32 pixel-responses). While one warp can process multiple lines, multiple warps within a thread block cooperatively process consecutive lines in a cell. As the corner response map is stored using a single-precision floating-point format in a pitched memory layout, the horizontal cell boundaries perfectly coincide with the L1-cache line boundaries, which maximizes the memory bus utilization whenever fetching complete cell line segments.

For the simplicity of illustration, a 1:1 warp-to-cell mapping is used in a 32x32 cell. As a warp reads out the first line of the cell, each thread within the warp, has acquired one pixel-response. As the next operation, the entire warp starts the neighborhood suppression: the warp starts spiraling according to [14], and each thread verifies whether the response it has is the maximum within its Moore neighborhood. Once

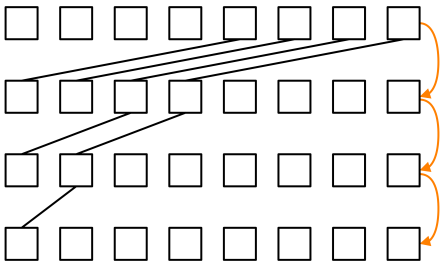


Fig. 2. Warp-level communication pattern during cell maximum selection. At the end of the communication, thread 0 has the valid maximum.

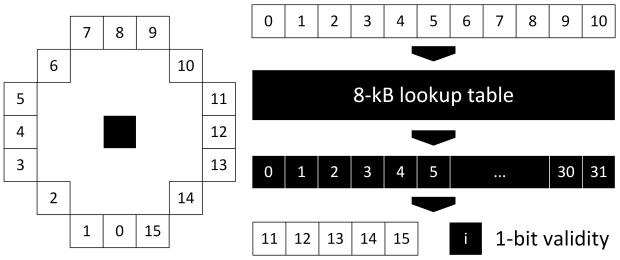


Fig. 3. FAST corner point evaluation with an 8 kilobyte lookup table

the neighborhood verification finishes, a few threads might have suppressed their response. However, no write takes place at this point, every thread stores its state (response score, and x-y location) in registers. The warp continues with the next line, and repeats the previous operations: they read out the corresponding response, and start the neighborhood suppression, and update their maximum response if the new response is higher than the one in the previous line. The warp continues this operation throughout all cell lines until it processed the entire cell. Upon finishing, each thread has its maximum score with the corresponding 2D location. However, as the 32 threads were processing individual columns, the maximum is only column-wise. Therefore, the warp needs to perform a warp-level reduction to get the cell-wise maximum: they reduce the maximum score and location to the first thread (thread 0) using warp-level shuffle down reduction [32]. The applied communication pattern is shown in Figure 2. Thread 0 finally writes the result to global memory.

To speed up reduction, multiple (M) warps process one cell, therefore, after the warp-level reduction, the maximum is reduced in shared memory. Once all warps within the thread block wrote their maximum results (score, x-y location) to their designated shared memory area, the first thread in the block selects the maximum for each cell and writes it to global memory, finishing the processing of this cell.

During pyramidical feature detection, we maintain only one grid. On level 0 (original resolution), the above-specified algorithm applies. On lower pyramid levels, we virtually scale the cell sizes, such that the applicable cell size on level k becomes $(\frac{32 \cdot w}{k}, \frac{2^{l-1}h}{k})$. In case the cell width falls below 32, one warp may process multiple lines: if the consecutive cell lines still belong to the same grid cell, the warp can analogously perform the warp-level reduction. Since a lower pyramid level's resolution is half of its upper layer's resolution, we can efficiently recompute where a pixel response falls from lower pyramid levels on the original grid. That is, when we identify a cell maximum on a lower pyramid level, the 2D position (x, y) from the lower resolution can be scaled up to $(2^l x, 2^l y)$.

Looking back at Algorithm 1, our approach combines the regular neighborhood suppression (NMS) and cell maximum selection (NMS-C) into a single step. It also differs from [11], [12], because we first perform candidate suppression within each thread in parallel, then reduce the remaining candidates amongst one cell.

D. FAST feature detector

The FAST feature detector's underlying idea is simple: for every pixel location in the original image (excluding a minimum border of 3 pixels on all sides) we perform a segment test, in which we compare pixel intensities on a Bresenham circle with a radius of 3. This Bresenham circle gives us 16 pixel-locations around each point (see Figure 3). We give labels L_x to these points based on a comparison between the center's and the actual point's intensity.

Given there is a continuous arc of at least N pixels that are labeled either *darker* or *brighter*, the center is considered a corner point. To add more robustness to the comparisons, a threshold value (ϵ) is also applied. The comparisons are summarized in (1). Both the number of continuous pixels (N) and the threshold value (ϵ) are tuning parameters.

$$L_x = \begin{cases} \text{darker} & I_x < I_{center} - \epsilon \\ \text{similar} & I_{center} - \epsilon \leq I_x \leq I_{center} + \epsilon \\ \text{brighter} & I_{center} + \epsilon < I_x \end{cases} \quad (1)$$

Avoiding Execution Divergence in FAST Calculation

If each thread performed the comparisons from (1) in NVIDIA's single-instruction-multiple-threads (SIMT) execution model, the comparison decisions in *if/else*-instructions will execute different code blocks. Since all threads execute the same instruction in a warp, some threads will be inactive during the *if*-branch and others during the *else*-branch. This is called code divergence and reduces the throughput in parallelization significantly, but it can be resolved with a completely different approach: a lookup table (Figure 3).

Our approach stores the result of the 16 comparisons as a bit array, which serves as an index for the lookup table. All possible 16-bit vectors are precalculated: a bit b_x is '1' if the pixel intensity on the Bresenham circle I_x is darker/brighter than the center pixel intensity I_{center}, and '0' if the pixel intensities are similar. As the result is binary for all 2^{16} vectors, the answers can be stored on 2^{16} bits, i.e. 8 kilobytes. These answers can be stored by using 4-byte integers, each of which store 32 combinations (2^5): 11 bits are used to acquire the address of the integer, and the 5 unused bits then select one bit out of the 32. If the resulting bit is set, we proceed with the calculation of the corner score.

The literature distinguishes three different types of scores for a corner point: sum of absolute differences on the entire

Bresenham circle (SAD-B); sum of absolute differences on the continuous arc (SAD-A) [3]; maximum threshold (ϵ) for which the point is still considered a corner point (MT) [7]. The corner score is 0 if the segment test fails.

Our approach compresses the validity of each 16-bit combination into a single bit, resulting in an 8-kilobyte lookup table, for which the cache-hit ratio is higher than the work presented in [10]. This results from the increased reuse of each cache line that is moved into the L1 (and L2) caches, therefore improving the access latency.

E. Lucas-Kanade Feature Tracker

Our approach deploys the pyramidical approximated simultaneous inverse compositional Lucas-Kanade algorithm as feature tracker. The Lucas-Kanade [19] algorithm minimizes the photometric error between a rectangular patch on a template and a new image by applying a warping function on the image coordinates of the new image. The inverse compositional algorithm is an extension that improves on the computational complexity per iteration [20] by allowing to precompute the Hessian matrix and reuse it in every iteration. The simultaneous inverse compositional Lucas-Kanade adds the estimation of affine illumination change. However, as the Hessian becomes the function of the appearance estimates, it cannot be precomputed anymore, which makes this approach even slower than the original Lucas-Kanade. Therefore, our approach applies the approximated version, where the appearance parameters are assumed to not change significantly, and hence the Hessian can be precomputed with their initial estimates [21].

We use a translational displacement model $\boldsymbol{t}$ with affine intensity variation estimation $\boldsymbol{\lambda}$. The complete set of parameters are $\boldsymbol{q} = [\boldsymbol{t}, \boldsymbol{\lambda}]^\intercal = [t_x, t_y, \alpha, \beta]^\intercal$, where t_x, t_y are the translational offsets, while α, β are the affine illumination parameters, resulting in the warping

$$\boldsymbol{W}(\boldsymbol{x}, \boldsymbol{t}) = \begin{pmatrix} x + t_x \\ y + t_y \end{pmatrix}. \tag{2}$$

The per-feature photometric error that we try to minimize for each feature with respect to $\Delta\boldsymbol{q} = [\Delta\boldsymbol{t}, \Delta\boldsymbol{\lambda}]$ is

$$\begin{aligned} \min \sum_{\boldsymbol{x} \in N} \Big[& T(\boldsymbol{W}(\boldsymbol{x}, \Delta\boldsymbol{t})) - I(\boldsymbol{W}(\boldsymbol{x}, \boldsymbol{t})) + \\ & (\alpha + \Delta\alpha) \cdot T(\boldsymbol{W}(\boldsymbol{x}, \Delta\boldsymbol{t})) + (\beta + \Delta\beta) \Big]^2, \end{aligned} \tag{3}$$

where $T(\boldsymbol{x})$ and $I(\boldsymbol{x})$ stand for the template image and the current image intensities at position $\boldsymbol{x}$, respectively. The vector $\boldsymbol{x}$ iterates through one feature's rectangular neighborhood ($\mathcal{N}$). We can organize the coefficients of the incremental terms into vector form as

$$\boldsymbol{U}(\boldsymbol{x}) = \begin{bmatrix} (1+\alpha)\frac{\partial T(\boldsymbol{x})}{\partial x}\frac{\partial \boldsymbol{W}(\boldsymbol{x},\boldsymbol{t})}{\partial t_x} \\ (1+\alpha)\frac{\partial T(\boldsymbol{x})}{\partial y}\frac{\partial \boldsymbol{W}(\boldsymbol{x},\boldsymbol{t})}{\partial t_y} \\ T(\boldsymbol{x}) \\ 1 \end{bmatrix}, \tag{4}$$

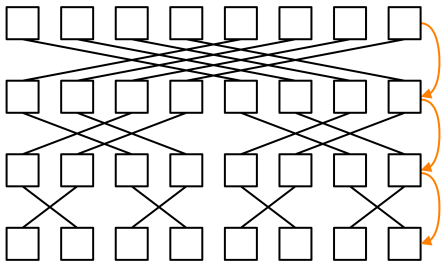


Fig. 4. Warp-level communication pattern during each minimization iteration, after which each thread has the sum of all individual thread results.

and the minimization problem can be rewritten as

$$\min \sum_{\boldsymbol{x} \in N} \Big[(1+\alpha)T(\boldsymbol{x}) + \beta - I(\boldsymbol{W}(\boldsymbol{x}, \boldsymbol{t})) + \boldsymbol{U}^\intercal(\boldsymbol{x})\Delta\boldsymbol{q} \Big]^2. \tag{5}$$

After computing the derivative of (5) and setting it to zero, the solution to $\Delta\boldsymbol{q}$ is found using the Hessian $\boldsymbol{H}$ by

$$\begin{aligned} \Delta\boldsymbol{q} &= \boldsymbol{H}^{-1} \sum_{\boldsymbol{x} \in N} \boldsymbol{U}^\intercal(\boldsymbol{x}) \left[I(\boldsymbol{W}(\boldsymbol{x}, \boldsymbol{t})) - (1+\alpha)T(\boldsymbol{x}) - \beta \right] \\ \boldsymbol{H}(\boldsymbol{x}) &= \sum_{\boldsymbol{x} \in N} \boldsymbol{U}^\intercal(\boldsymbol{x}) \boldsymbol{U}(\boldsymbol{x}). \end{aligned} \tag{6}$$

For this algorithm, there are two GPU problems that need to be addressed: memory coalescing and warp divergence. The problems are approached from the viewpoint of VIO algorithms, where one generally does not track a high number of features (only 50-200), and these sparse features are also scattered throughout the image, which means that they are scattered in memory.

This algorithm minimizes the photometric error in a rectangular neighborhood around each feature on multiple pyramid levels. Consequently, if threads within a warp processed different features, the memory accesses would be uncoalesced, and given some feature tracks do not converge or the number of iterations on the same level differs, some threads within a warp would be idle. To address both of these concerns, one entire warp is launched for processing one feature. We also opted for rectangular patch sizes that can be collaboratively processed by warps: on higher resolutions 16x16, on lower resolutions 8x8 pixels. It solves warp divergence since threads within the warp perform the same number of iterations and they iterate until the same pyramid level. The memory requests from the warp are also split into fewer memory transactions, as adjacent threads are processing consecutive pixels or consecutive lines.

Warp-level primitives are exploited in every iteration of the minimization, as we need to perform a patch-wide reduction: in (6) we need to sum a four element vector $\boldsymbol{U}(\boldsymbol{x})^T r(\boldsymbol{x}, \boldsymbol{t})$ for every pixel within the patch. One thread processes multiple pixels within the patch, hence each thread reduces multiple elements into its registers prior to any communication. Once the entire warp finishes a patch, the threads need to share their local results with the rest of the warp to calculate $\Delta\boldsymbol{q}$. This reduction is performed using the butterfly communication pattern shown in Figure 4.

TABLE II
COMPARISON OF GPU EVALUATION HARDWARE

	Tegra X2	960M
GPU Tier	embedded	notebook
CUDA Capability	6.2	5.0
CUDA Cores	256	640
Maximum GPU Clock	1300 MHz	1176 MHz
Single-precision Performance	665.6 GF/s	1505.28 GF/s
Memory Bandwidth	59.7 GB/s	80 GB/s

The novelty of our approach lies in the thread-to-feature assignment. Approaches presented in [25], [26] are using a one-to-one assignment, which implies that only large feature-counts can utilize a large number of threads. This adversely affects latency hiding on GPUs with smaller feature counts, which is generally applicable to VIO. Our method speeds up the algorithm by having warps that collaboratively solve feature patches, where each thread's workload is reduced, while the used communication medium is the fastest possible.

III. EVALUATION

We evaluate in four parts: non-maxima suppression, standalone feature detection, feature tracking (all on the EuRoC Machine Hall 01 sequence [27], including 3,682 image frames), and applicability within a VIO pipeline. The full VIO pipeline is implemented with the bundle-adjustment from [33] and tested on the Machine Hall EuRoC dataset sequences [27].

A. Hardware

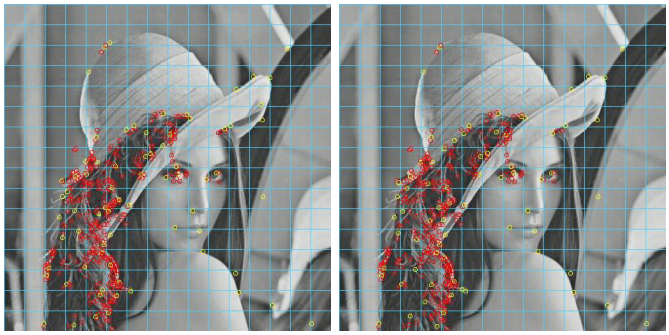
We performed our experiments on an NVIDIA Jetson TX2 and a laptop computer with an Intel i7-6700HQ processor and a dedicated NVIDIA 960M graphics card. The Jetson TX2 was chosen because of its excellent tradeoff between size, weight, and computational capabilities, where we run all experiments with the platform in *max-N* performance mode (all cores and GPU at maximal clock speeds). The properties of the two platforms are summarized in Table II.

B. Non-Maxima Suppression

The proposed FAST corner response function ($\epsilon = 10$, $N = 10$) was used as input to our non-maxima suppression, run with a grid granularity of 32×32 ($l = 1, w = 1, h = 32$) and compared to Förstner's [14] spiral non-maxima suppression algorithm. The results are listed in Table III, showing a 2× speedup for the embedded Jetson TX2 platform.

C. Feature detector

1) Conformance: We verify our feature detection conformance with the original FAST feature detector, for both suggested score functions: 5(a) the sum of the absolute difference between the center pixel and the contiguous arc; 5(b) the maximum threshold value, for which the point is detected as a corner. As our combined non-maxim suppression selects a single maximum within each cell, our output only comprises of a subset of the original implementation's output. In Figure 5 we mark features red ○ which are output from the original detector, yellow ○ which are output from both implementations, and blue ○ false-positives of our detector. Note that there are no false-positives and that our method returns a well-distributed subset of the original response, rendering additional feature selection unneccessary,



(a) Sum of absolute differences [3] (b) Maximum threshold value [7]

Fig. 5. Conformance verification with the original FAST detector (○) and our combined FAST/NMS (○ for conforming detections, ○ for false-positives). Note that there are no false-positives and that our method returns a well-distributed subset of the original response, therefore rendering additional feature selection unnecessary. Best viewed in digital paper.

2) Cache-hit ratio: As mentioned in II-D, we expect higher GPU cache-hit ratios during feature detection with our bit-based CRF lookup table. This reduces the number of global-memory transactions, resulting in lower kernel execution times. The cache-hit ratios and the resulting CRF timings are listed in Table IV.

3) Execution time breakdown: We split the execution time of pyramidal feature detection ($l = 2$) into its constituents: image copy from host to device memory (*Upload*), creation of an image pyramid where each subsequent layer halves the resolution of the previous one (*Pyramid*), corner response function evaluation (*CRF*), non-maxima suppression with cell-maximum selection (*NMS*), and feature grid copy from device memory to host memory (*Download*).

4) Execution time comparison: We compare our feature detection with other publicly available FAST implementations, summarized in Table V. As the publicly available detectors only support single-scale, we performed these experiments

TABLE III
COMPARISON OF 2D NON-MAXIMA SUPPRESSION KERNELS ON GPU

	Tegra X2		960M	
NMS method	Förstner	Ours	Förstner	Ours
n=1 (3x3)	294.03 μs	**141.36 μs**	96.81 μs	**73.78 μs**
n=2 (5x5)	696.57 μs	**338.03 μs**	245.96 μs	**158.93 μs**
n=3 (7x7)	1207 μs	**604.94 μs**	441.59 μs	**288.39 μs**
n=4 (9x9)	1772 μs	**929.82 μs**	661.90 μs	**450.08 μs**

TABLE IV
TIMING COMPARISON OF DIFFERENT FAST CRF SCORES

		Tegra X2		960M	
Lookup table		byte-based	bit-based	byte-based	bit-based
SAD-B	**L1 cache-hit rate**	83.9 %	**89.9 %**	68.6 %	**77.4 %**
	CRF kernel	317.3 μs	**298.7 μs**	141.3 μs	**135.9 μs**
SAD-A	**L1 cache-hit rate**	83.9 %	**89.8 %**	68.5 %	**77.3 %**
	CRF kernel	348.4 μs	**334.9 μs**	158.5 μs	**155.1 μs**
MT	**L1 cache-hit rate**	84.0 %	**91.8 %**	71.9 %	**82.7 %**
	CRF kernel	815.1 μs	**784.3 μs**	410.6 μs	**374.8 μs**

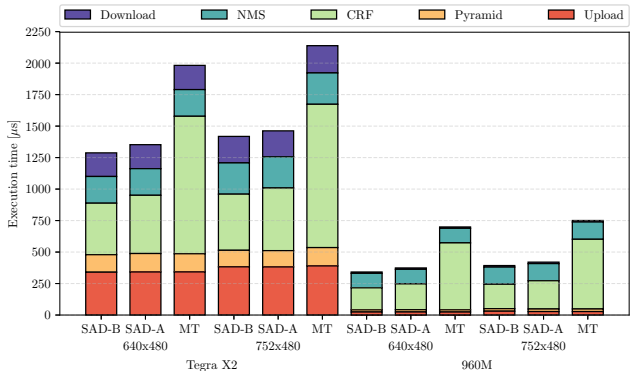


Fig. 6. Feature detector execution time breakdown into image upload, pyramid creation, CRF, NMS, and feature download. Best viewed in color.

on the original image resolution. Note that our method not only performs feature extraction but simultaneously applies cell-wise non-maxima suppression, still achieving superior execution times. As KFAST uses x86-specific instruction set extensions, while ArrayFire OpenCL was incompatible with our available packages, these could not be run on the Jetson TX2. The GPU timings include the image upload and feature list download times as well.

D. Feature tracker

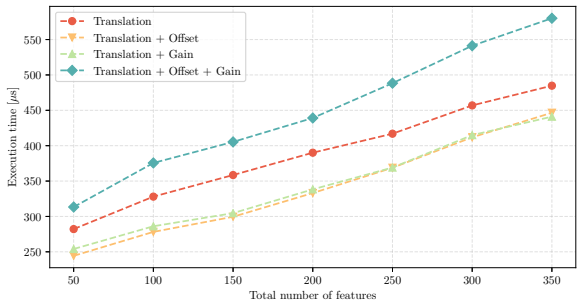
We timed our feature tracker implementation by varying the total number of simultaneous feature tracks on both testing platforms, depicted in Figure 7. We utilized our FAST with score (ii) as feature detector, and triggered feature re-detection whenever the feature track count falls below 30% of the actual target. The evaluations include the full affine intensity and translation estimation, with a total of 4 estimated parameters. We also show the translation-only estimation, translation-gain, as well as translation-offset estimation. Note that translation-only estimation is less efficient since it needs more iterations to converge, as visible in Figure. 7. Lastly, Table VI shows a comparison against OpenCV's CPU and GPU implementation, which we outperform by a factor of 2× and more.

E. Visual odometry

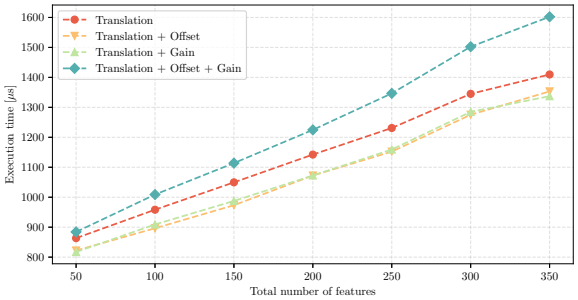
Lastly, we evaluate the performance of our frontend in combination with a VIO bundle-adjustment backend. We chose ICE-BA [33] as backend, since it is accurate, efficient and also achieves extremely fast execution times. We implement two

TABLE V
FAST FEATURE DETECTOR AVERAGE EXECUTION TIME COMPARISON

	Tegra X2	**960M**
Others (feature detection and NMS)		
OpenCV CPU with MT	4.78 ms	2.23 ms
OpenCV CUDA with MT	2.73 ms	1.09 ms
ArrayFire CPU with SAD-A	60.83 ms	36.42 ms
ArrayFire CUDA with SAD-A	1.47 ms	0.51 ms
ArrayFire OpenCL with SAD-A	-	0.91 ms
KFAST CPU with MT	-	0.63 ms
Ours (feature detection, NMS, and NMS-C)		
Ours with SAD-B	1.11 ms	0.28 ms
Ours with SAD-A	1.14 ms	0.30 ms
Ours with MT	1.59 ms	0.52 ms



(a) Performance on 960M



(b) Performance on Tegra X2

Fig. 7. Feature tracker performance comparison over a varying number of tracked features, with translation, illumination gain and offset estimation, and subsets of those. Best viewed in color.

test cases: first, we run the ICE-BA backend with their proposed frontend as a baseline and then compare it against our frontend with their backend. Both frontends employ FAST features (our implementation vs. theirs), a Lucas-Kanade feature tracker with 70 tracked features. We tuned the original ICE-BA configuration [33] and reduced the local bundle adjustment (LBA) window size to 15 frames for both cases, while we kept other parameters unchanged. We summarize our findings in Table. VII, which shows an average speedup factor of 2.25× of the combined front- and back-end on both platforms, with an average accuracy loss of 0.47%. Our tracking performance on MH_04 and MH_05 is affected by the faster motions and dark scenes. However, according to [34], most pipelines suffer from largely increased tracking error on these two sequences. This is due to the relatively dark appearance combined with fast motions of some scenes in these sequences, introducing higher noise in feature tracking. The VIO pipeline combining our GPU-accelerated frontend and the CPU targetted ICE-BA backend allows us to achieve a throughput of ∼200fps on multiple datasets on the embedded Jetson TX2 platform.

TABLE VI
FEATURE TRACKER AVERAGE EXECUTION TIME COMPARISON, TRACKING 100 FEATURES, RE-DETECTION AT 30 FEATURES

	Tegra X2	**i7-6700HQ+960M**
OpenCV CPU	1.88 ms	1.38 ms
OpenCV CUDA	3.96 ms	0.74 ms
Ours trans. only	0.96 ms	0.33 ms
Ours trans. & offset	0.90 ms	0.28 ms
Ours trans. & gain	0.91 ms	0.29 ms
Ours trans. & gain & offset	1.01 ms	0.38 ms

TABLE VII
RESULTS OF OUR FRONTEND COMBINED WITH A VIO BACKEND [33] ACHIEVING ∼200FPS THROUGHPUT ON THE EUROC DATASET [27]

		Average execution time					Relative translation error (RMSE)				
		MH_01	MH_02	MH_03	MH_04	MH_05	MH_01	MH_02	MH_03	MH_04	MH_05
Tegra X2	**Original**	11.64 ms	12.90 ms	12.91 ms	12.90 ms	12.85 ms	1.08 %	0.71 %	0.40 %	0.81 %	0.50 %
	Ours	4.67 ms	4.93 ms	6.41 ms	6.10 ms	5.87 ms	0.86 %	0.90 %	1.40 %	1.85 %	1.19 %
i7-6700HQ+960M	**Original**	4.11 ms	4.28 ms	4.77 ms	4.63 ms	4.64 ms	0.68 %	0.71 %	0.53 %	0.83 %	1.36 %
	Ours	1.56 ms	1.74 ms	2.54 ms	2.51 ms	2.06 ms	1.00 %	0.55 %	0.74 %	2.14 %	1.68 %

IV. CONCLUSION

This work introduces a novel non-maxima suppression exploiting low-level GPU-specific instruction primitives, complemented by a GPU-accelerated FAST feature detector implementing multiple corner response functions. Our approach is unique in the way it combines feature detection and non-maxima suppression, not only guaranteeing uniform feature distribution over an image but also consistently outperforming all other available implementations in terms of execution speed. The speed improvement for the embedded computer is more pronounced, as there's a higher performance penalty for memory interactions than in the case of the laptop GPU. We verified the conformity with the original FAST detector, analyzed the execution timings on two different platforms considering corner response functions, non-maxima suppression, and feature tracking on multiple numbers of features. As opposed to others, our feature tracker utilizes a feature-to-warp assignment, which speeds up tracking operations in typical VIO scenarios. Finally, we demonstrate superior speed in combining our frontend with a VIO bundle-adjustment backend, achieving a metric state estimation throughput of ∼200 frames per second with high accuracy on an embedded Jetson TX2 platform, providing a real-time, heterogeneous VIO alternative.

REFERENCES

[1] C. Harris and M. Stephens. A combined corner and edge detector. In *In Proc. of Fourth Alvey Vision Conference*, 1988.

[2] Jianbo S. and Carlo T. Good features to track. *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 1994.

[3] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *Eur. Conf. Comput. Vis. (ECCV)*, 2006.

[4] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *Int. Conf. Comput. Vis. (ICCV)*, 2011.

[5] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 2004.

[6] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image. Und.*, 2008.

[7] E. Rosten, R. Porter, and T. Drummond. Faster and better: A machine learning approach to corner detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010.

[8] K. Omar. Kfast: vectorized x86 cpu implementation of the fast feature detector, 2006.

[9] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[10] P. Yalamanchili, U. Arshad, Z. Mohammed, P. Garigipati, P. Entschev, B. Kloppenborg, J. Malcolm, and J. Melonakos. ArrayFire - A high performance software library for parallel computing with an easy-to-use API, 2015.

[11] A. Neubeck and L. Van Gool. Efficient non-maximum suppression. In *IEEE Int. Conf. Pattern Recog. (ICPR)*, 2006.

[12] Tuan Q. Pham. Non-maximum suppression using fewer than two comparisons per pixel. In *Advanced Concepts for Intelligent Vision Systems ACIVS*, 2010.

[13] D. Oro, C. Fernández, X. Martorell, and J. Hernando. Work-efficient parallel non-maximum suppression for embedded gpu architectures. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[14] W. Förstner and E. Gülch. A fast operator for detection and precise location of distinct point, corners and centres of circular features. In *Proceedings of the ISPRS Conference on Fast Processing of Photogrammetric Data*, 1987.

[15] A. I. Mourikis and S. I. Roumeliotis. A multi-state constraint kalman filter for vision-aided inertial navigation. In *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2007.

[16] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale. Keyframe-based visual-inertial odometry using nonlinear optimization. *Int. J. Robot. Research*, 2015.

[17] Raúl Mur-Artal, José M. M. Montiel, and Juan D. Tardós. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Trans. Robot.*, 2015.

[18] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart. Robust visual inertial odometry using a direct EKF-based approach. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2015.

[19] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Int. Joint Conf. Artificial Intell. (IJCAI)*, 1981.

[20] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework: Part 1. *Int. J. Comput. Vis.*, 2002.

[21] S. Baker, R. Gross, and I. Matthews. Lucas-kanade 20 years on: A unifying framework: Part 3. *Int. J. Comput. Vis.*, 2003.

[22] C. Forster, M. Pizzoli, and D. Scaramuzza. Svo: Fast semi-direct monocular visual odometry. In *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2014.

[23] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza. SVO: semidirect visual odometry for monocular and multicamera systems. *IEEE Trans. Robot.*, 2017.

[24] C. Zach, D. Gallup, and J. Frahm. Fast gain-adaptive klt tracking on the gpu. In *IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, 2008.

[25] J.S. Kim, M. Hwangbo, and T. Kanade. Realtime affine-photometric klt feature tracker on gpu in cuda framework. In *Int. Conf. Comput. Vis. Workshops (ICCVW)*, 2009.

[26] M. Hwangbo, J. Kim, and T. Kanade. Inertial-aided klt feature tracking for a moving camera. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2009.

[27] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M.W. Achtelik, and R. Siegwart. The EuRoC micro aerial vehicle datasets. *Int. J. Robot. Research*, 2015.

[28] NVIDIA Corporation. *CUDA C++ Programming Guide*, 2019.

[29] D. Nister, O. Naroditsky, and J. Bergen. Visual odometry. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2004.

[30] D. Scaramuzza, F. Fraundorfer, and R. Siegwart. Real-time monocular visual odometry for on-road vehicles with 1-point ransac. In *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2009.

[31] F. Fraundorfer and D. Scaramuzza. Visual odometry : Part ii: Matching, robustness, optimization, and applications. *IEEE Robot. Autom. Mag.*, 2012.

[32] NVIDIA Corporation. *Developer Blog - Faster Parallel Reductions on Kepler*, 2014. https://devblogs.nvidia.com/faster-parallel-reductions-kepler/.

[33] H. Liu, M. Chen, G. Zhang, H. Bao, and Y. Bao. Ice-ba: Incremental, consistent and efficient bundle adjustment for visual-inertial slam. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018.

[34] Jeffrey Delmerico and Davide Scaramuzza. A benchmark comparison of monocular visual-inertial odometry algorithms for flying robots. In *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2018.